中文植物物种多样性描述文本的信息抽取研究*

段宇锋　黄思思
(华东师范大学商学院　上海 200241)

摘要:【目的】实现中文植物物种多样性描述文本中信息的抽取。【方法】以中文植物物种多样性本体为支撑, 采取语段、语句、概念逐级筛选和标注的策略, 依据规则抽取描述文本中的信息。【结果】以包含 4 734 个信息点的样本测试, 信息抽取的准确率、召回率、F 值分别为 0.86、0.85、0.85。【局限】针对目前未能准确抽取的表述, 进一步完善规则集。【结论】研究方案能有效地实现中文植物物种多样性描述文本的信息抽取。

关键词: 信息抽取　植物物种多样性描述文本　中文信息处理　本体
分类号: G350　TP18

物种是最接近生物的自然单元, 因此, 物种描述也就成为生物学和生态学研究的起点。过去的两百多年间, 人类在探索自然的过程中形成了海量的物种描述文献。以生物多样性历史文献库(Biodiversity Heritage Library, BHL)为例, 截至 2015 年 9 月 27 日, 其存储的文献已超过 4 727 万页[1]。这些文献的有效开发利用将极大地支持生物学和生态学的研究。从 20 世纪 80 年代的纸质文献数字化, 到现在的大规模网络共享, 无疑有效提升了这些文献的传播和利用效率。信息传递方式和渠道的变化, 使人们在获得丰富信息的同时, 也产生了巨大的筛选和处理压力。信息抽取技术的发展、成熟, 为问题的解决提供了思路和方法。

1　国内外研究现状

信息抽取就是识别和提取文档中用户感兴趣的内容, 并以结构化、语义清晰的形式表示。该领域起源于文本理解研究, 数字文本的急剧增长和消息理解系列会议(Message Understanding Conference, MUC)的推动, 使其逐步发展成为自然语言处理领域的一个重要分支。

生物物种多样性描述文本的信息抽取研究始于 20 世纪 90 年代中期。迄今, 虽然取得了一些进展, 但远未满足自动化地实现海量生物多样性信息的细粒度组织和语义检索的现实需求。Thessen 等将国外研究分为数字化(Digitization)、语义标注(Annotation)、命名识别(Names Recognition and Discovery)、形态特征提取(Morphological Character Extraction)四类, 并系统地进行综述[2]。依据粒度, 笔者将形态特征的提取研究分为语句和概念两类, 具体如下:

(1) 语句层的形态特征提取研究

物种描述具有基本一致的模式。以植物描述为例, 一般都是从生长习性、根、茎、叶、花、果实描述到物候学特征。对于较复杂的器官结构, 则依其构成进一步展开。譬如, 对叶的描述会细化至叶柄、叶片等部位。因此, 物种描述信息在整体上呈倒置的树形结构。正因为如此, 以语句为单位的物种形态特征提取可转化为逐层分类问题。

具体实现一般采用规则系统或统计学习方法, 当然, 也可以将两者结合起来构建综合性的算法。譬如, Vanel 在人工分析句法和词汇特征的基础上开发解析

通讯作者: 段宇锋, ORCID: 0000-0002-4319-2837, E-mail: yfduan@infor.ecnu.edu.cn。
*本文系国家社会科学基金一般项目"基于无监督语义标注的网络中文学术信息抽取研究"(项目编号:11BTQ024)的研究成果之一。

器，实现语句标注[3]。郑家恒等在聚类的基础上，利用主题分布的特点对农作物种子信息进行语句层标注[4]。Cui 等则依据物种描述文本的句子通常以表示植物结构的名词词组开头这一句法特征，将句子的先导词与词频分布相结合建立语句标注算法。以《Flora of North America》(FNA)和英文版《中国植物志》(FOC)中的文档为测试样本，标注的平均准确率和召回率都在0.9 以上[5]。本课题组与 Cui 合作，将该算法修正后应用于中文植物物种多样性描述文本的语句标注。以《中国植物志》中的文档作为测试样本，整体标注性能(F 值)达到 0.930[6]。为了降低标注系统的运行负荷，本课题组尝试将先导词与朴素贝叶斯统计学习方法相结合，其标注性能(F 值)也达到了 0.902[7]。上述研究虽然都获得了令人满意的标注结果，但都要耗费大量的专家资源，而且建立的规则和训练数据很难适应不同的文本集。鉴于此，笔者在朴素贝叶斯算法的基础上，引入 Bootstrapping 方法。采用与前两项研究相同的测试集检验算法性能，F 值为 0.9112，显著高于朴素贝叶斯与先导词相结合的算法($P<0.05$)。这一方法不仅极大地降低了系统对训练集规模和专家的依赖，而且有效提高了标注性能[8]。这也是本研究在语句标注阶段使用的算法。

(2) 概念层的形态特征提取研究

概念的语义理解是实现概念层形态特征提取的关键。因此，无论是依靠人工还是自动识别方式，所有研究都建立了与其目标相适应的术语集。在形式上，它可以表现为索引、词汇表甚至本体。这也同时决定了所有研究采用的都是基于规则的方法。

Taylor 在分析文本语法特征的基础上，以人工方式建立规则和词典，抽取《Flora of New South Wales》(第 4 卷)和《Flora of Australia》(第 19 卷)中的物种部位、特征和状态，召回率介于 0.6-0.8[9]。这是概念层物种描述信息抽取最早的研究。Wood 等依靠人工创建的领域本体和 GATE 提供的正则表达式匹配能力，实现植物描述特征的抽取，准确率、召回率为 74%和66%[10]。Tang 等改造 Soderland 提出的方法，依据有监督学习自动生成的规则，将北美植物群落 1 600 种物种的叶子的形状、大小、颜色、排列及果实的形状特征填充到预先定义的模板，准确率介于 30%-100%[11-12]。Abascal 等、Diederich 等将人机交互引入特征抽取过程，分别建立了 X-Tract、Terminator 系统。在实现原理上，两者与上述研究相同[13-14]。Cui 等采用启发式方法和句法特征生成规则，从 FNA 第 19 卷和《Treatise on Invertebrate Paleontology》(TIP)H 部分分别取 400 篇文档进行测试。前者在两个文本集中抽取的准确率和召回率分别是 0.63、0.6 和 0.52、0.43，后者为 0.91、0.9 和 0.8、0.87[15-16]。

由于中文在构词、句法等方面与英文差异显著，所以，国外的研究成果基本无法直接应用于中文物种描述文本的信息抽取。迄今，国内与本项目相似的研究只有两项。其一，沙丽华依靠建立的玉米本体标注文档中的概念、属性和实例，并以三元组表示[17]，该研究与本项目的整体思路比较相似，但处理的并非物种多样性描述文本，且仅涉及玉米领域；其二，石静在植物本体概念系统的支持下，标注植物描述文本中出现的概念和实例，实现句子分类，进而据此选择抽取模板并依据定义的规则和标注结果抽取实体填充模板[18]，该研究采用固定模板和人工构建的规则，通常会面临灵活性、适应性的问题。

本文以实现中文植物物种多样性描述文本中信息的抽取为目标，希望建立的方案既能用于单一器官结构特征的抽取，也能支持全文本信息抽取，且在不同文本集具有良好适应性。因此，关键在于最大限度地识别和标注领域概念，这是选择基于本体的方法最主要的原因；而且，随着本体概念体系的完善，系统的抽取性能和适应性将不断提升，这是采用基于本体的方法的另一个原因。

2　信息抽取方案

本研究将中文植物物种多样性描述性文本的信息抽取分解为 4 项任务：构建领域本体、建立训练集数据、文本预处理、标注和抽取，如图 1 所示。

2.1　构建中文植物物种多样性领域本体

本体是概念模型的明确的规范说明和定义[19]。领域本体提供了特定领域中概念和关系的描述。在研究方案中，领域本体主要有三方面的作用：解析本体中的概念，生成领域词典，从而提高系统分词的精确性；将概念的语义类作为 CRF 算法的特征，识别待处理文本中的新概念；支持文本标注，并依据本体建立所标注概念之间的关系，实现信息抽取。本研究以 BFO 为

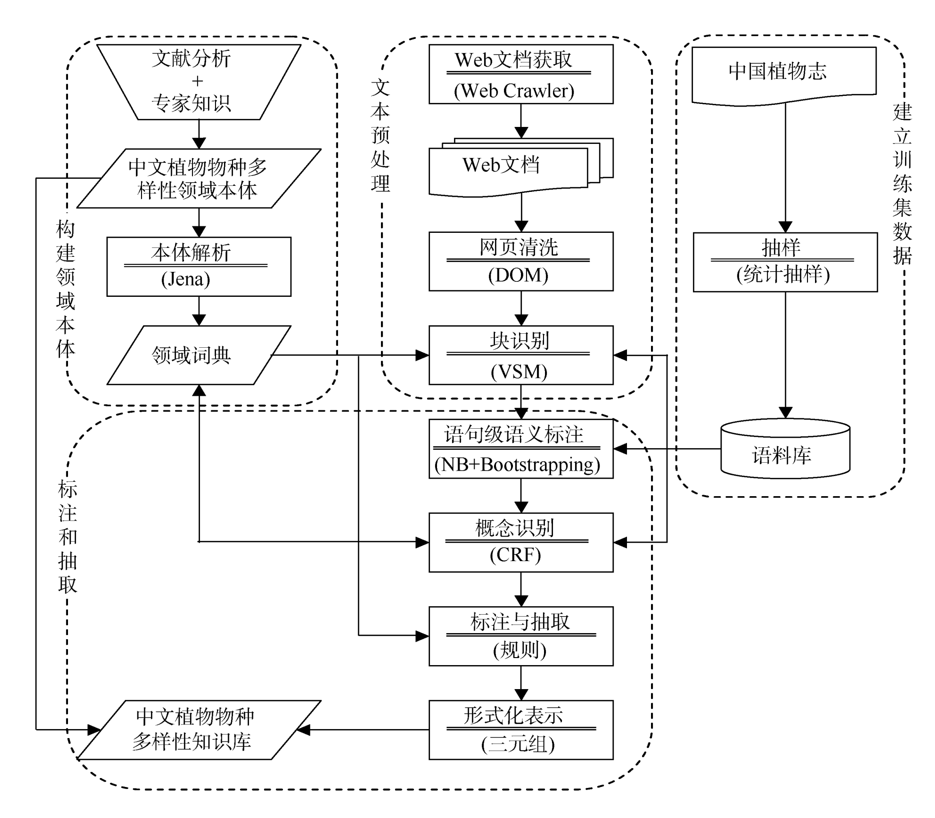


图 1 研究思路和方法

上层本体，采用 KACTUS 法复用 PO，建立的中文植物物种多样性本体含有 720 个类，4 000 多个实例。

(1) 类

①复用 PO 的类。保留 PO 的以下类项：植物结构下的复合植物结构的基本部分、毛状体、多组织植物结构和复合植物结构；形成空间的植物解剖结构；果实生长阶段；复合植物器官生长阶段下的花的生长阶段。并且，合并整理如下类项：将植物结构下保留的部分及“形成空间的植物解剖结构”合并为类“植物解剖结构”；将果实生长阶段、复合植物器官生长阶段下的花的生长阶段合并成为类“植物生长阶段”。同时，对复用的类添加对应的中文术语，并使其成为主要描述。

②增添类。增加植物分类、植物空间部位、物种和部位的属性、度量单位、程度限定等类。

(2) 实 例

PO 中只有类和关系，没有实例。本研究构建本体的目的是支持植物物种多样性描述文本中知识的抽取。前述表征植物物种多样性特征的属性类，如颜色、形状、质地等，不包含具体的实例。在缺乏实例的情况下，无法有效地抽取植物物种多样性信息。例如，在“花瓣 5，白色或淡红色”这句描述中，“白色”和“淡红色”是颜色的实例。如果实体“颜色”未包含这两个实例，就难以判定该句是描述花的颜色，也就无法准确提取相应的信息。

实例的数量众多。在依靠领域专家定义的基础上，本研究还采用了自动识别技术，在文档处理过程中发现和填充新实例。具体方式为：

①采用条件随机场(CRFs)算法识别候选实例；

②人工筛选候选实例，确定新增实例列表；

③逐行读取列表，向本体文件中添加相应实例描述语句。譬如，增加形状实例“长圆状卵形”时向本体文件中添加语句：

```
<!-- http://www.ontology/plant-species-diversity#长圆状卵形 -->
    <owl:NamedIndividual rdf:about="&psd;长圆状卵形">
        <rdf:type rdf:resource="&psd;形状"/>
    </owl:NamedIndividual>
```

(3) 关 系

①复用 PO 的关系，包括：adjacent_to、derives_by_ manipulation_from、developmentally_preceded_by、part_of、has_ part、develops_from、has_participant、located_in、participates_ in、preceded_by。

②增添关系，具体如下：

1)特征关系。用于呈现植物结构与相关特征的关联。包

括 has_color、has_shape、has_arrangement、has_texture、has_growth_form、has_accy_structure、distributed_in、has_flower_period、has_fruit_period、has_height、has_weight、has_length、has_diameter、has_quantity。

2)分类关系。用于揭示植物分类知识。包括 has_class、has_order、has_family、has_gensus、has_species。

3)附属关系。用于完善相关特征描述。包括 has_unit、has_degree。

2.2 建立训练集

Web 文本的规范化程度低、文档集之间的差异大，训练数据的代表性对性能具有重要影响。以兼顾性能和通用性为原则，本研究采取随机抽样和分层抽样相结合的方式，从权威数据源《中国植物志》中获取 1 000 个物种的描述文本，共涉及 37 个科，每科大约 30 个种。每个物种的描述都独立地以 TXT 文档形式存储。

在研究方案中，文本预处理任务的块识别过程、信息抽取模块的语句级标注和概念识别过程分别采用不同的机器学习方法，需要建立相应的训练数据。具体如下:

(1) 支持块识别的训练数据

文本预处理中的块识别采用向量空间模型(Vector Space Model, VSM)，可以直接以 TXT 文档的内容作为训练数据。

(2) 支持语句级标注的训练数据

语句级标注需要解析到在句法上完整的句子，因此，对 1 000 个 TXT 文档中的内容以“;”和“。”为标识，以人工方式逐句标注。使用的语义标签包括“plant-habit-and-life-style”、“roots”、“stems”、“buds”、“leaves”、“flowers”、“fruits”、“seeds”、“spore-related-structures”、“phenology”和“compound”。其中，“compound”用于标注描述了两种或两种以上植物结构的语料，例如“苞片和小苞片线形”。每个 TXT 文档对应一个标注后形成的同名 XML 文档，该文档集即为语句级标注的训练数据。

(3) 支持概念识别的训练数据

概念识别过程采用 CRF 算法，以字、词为处理单元。为提高识别性能，本研究依据语句级标注训练数据含有的语义标签(“compound”除外)，构建相应的训练数据文档(TXT 格式)。训练数据以词、词性、词长、相关度、信息熵为特征，采用 SBIEO 作为标注集(见 2.4 节中的“(4)概念识别”)。

2.3 文本预处理

使用爬虫从网上获取文档。由于这些文档的格式、结构、编码方式可能各不相同，因而需要进行规范化处理，并筛选出与主题相关的文本块，传递给信息抽取模块。

(1) 网页清洗

网页是使用标记语言构建的半结构化文本。将网页解析成 DOM 树，去除与主题无关的<script>、<link>、<img>、<style>等元素，提取文本节点的内容并进行规范化处理，包括转换编码方式、剔除乱码和空格、将半角的标点符号转化为全角。

(2) 块识别

并非网页的所有文本节点都与描述内容相关，因此，本研究采用向量空间模型，以 0.8 为阈值，筛选文本节点内容。

2.4 信息抽取

信息抽取的基础是计算机能够理解构成自然语言文本的字符(串)的语义以及相互间的语义关系，因而，概念标注和关系识别无疑是实现抽取的关键。因为领域本体涵括了对概念和关系的描述，所以，本研究将本体作为实现信息抽取的关键支撑要素。并且，方案采用了从语句到字符逐级细化的标注过程，以提高标注的准确性，进而达到提升抽取性能的目的。

(1) 语句标注

语句标注采用与 Bootstrapping 方法相结合的朴素贝叶斯算法。实验结果表明，种子数达到 90 时，该方法的标注性能就已超过依靠大规模人工训练集支持的朴素贝叶斯算法[7]。而建立样本量仅为 90 的训练数据，耗费的时间和专家资源几乎可以忽略不计。

(2) 概念标注

在领域词典的支持下，调用 ICTCLAS①实现分词并添加词性或语义标识。语句标注为概念标注提供两方面的支持：一是验证分词的正确性，尤其是表示描

①http://ictclas.nlpir.org.

述主体的概念；二是语句缺省表示描述主体的概念时，将语句标注结果作为补充的依据。概念标注需要领域词典的支持，并事先定义标注集。

①生成领域词典

Jena 是基于 Java 开发的开放源代码语义网工具，提供了面向本体的模型处理、解析查询、基于规则的推理、持续性存储、不同本体形式的接口支持等多种功能[20]。其中，解析模块具有大量支持对元素进行操作的函数。本研究利用 listClasses()、listObjectProperties()、listDatatypeProperties()、listSuperClasses()、getDomain()等函数解析本体。以"苞腋"为例，Jena 的输出为：

类 URI: http://purl.obolibrary.org/obo/PO_0025225
类名: http://purl.obolibrary.org/obo/PO_0025225
标签：苞腋
类描述类型: subClassOf
类描述值: http://purl.obolibrary.org/obo/PO_0025224(枝腋)
类描述类型: subClassOf
类描述值：植物构成
类描述类型: subClassOf
类描述值: http://purl.obolibrary.org/obo/PO_0025131(植物结构)
类描述类型: subClassOf
类描述值: http://purl.obolibrary.org/obo/PO_0025117(珠孔)

注：本研究构建的本体复用了 PO；"()"是为便于理解解析结果而添加的注释。

从上述解析结果中提取类和实例的相应信息构建领域词典，词条格式为"XX instance/class 'class'"。其中，XX 为概念的标签；instance/class 用于表示概念的类型；'class'则表明该概念所属类。如概念为类，此处标识与 XX 相同。譬如，上例中的"苞腋"对应的条目形式为"苞腋 class 苞腋"。

②标注集

标注集是表示词汇语义的标识集合，用以标记分词文件中词汇的语义信息。标注集中大部分标识的涵义对应于本体第三层的概念和相应属性。此外，还有少量标识与本体中的概念无关，但与物种特征的描述密切相关。譬如，"密被"和"疏被"难以纳入本体的概念系统，但却常常出现于物种解剖结构的特征描述之中。依据信息抽取的需要，本研究设定以下标识，如表 1 所示。

表 1 概念标注集

标识	涵义
jg	植物解剖结构
ys	颜色
xz	形状
xt	形态
zd	质地
szx	生长型
pbei	用于描述植物结构上生长有其他附属结构的特定连接词，如"密被"、"疏被"等
kj	植物空间部位
hq	花期
gq	果期
dm	地域名称
cd	程度限定
dw	度量单位
n	会出现在特定植物部位上但并不存在于本体内的特定对象，如"点"、"网格纹"等

标注格式为"'标识':class/ins-'class'"。与词对应的概念若在本体中为类，":"后使用"class"；若为实例，":"后则使用"ins-"加其所属类名。由于标注在分词过程中同步进行，类名不宜使用中文形式，因此，实例所属类以英文或 OBO 编号表示。领域词典之外的字词和符号则保留分词时标注的词性标识。以勾儿茶属勾儿茶种的描述(部分)为例，标注结果如下所示：

藤状/szx:ins-growth_form 或/c 攀援灌木/szx:ins-growth_form ，/wd 高 /xt:ins-arrangement 达 /v 5/m 米 /dw:ins-unit ：/wp 幼枝 /jg:ins-PO_0025073 无/v 毛/jg: ins-PO_0000282 ，/wd 老枝/jg:ins-PO_0025073 黄褐色/ys:ins-color ，/wd 平滑/zd:ins-texture 无/v 毛 /jg: ins-PO_0000282 。/wj

(3) 抽取规则和抽取过程

①抽取规则

一方面，本研究试图建立具有广泛适用性的描述性文本信息抽取方案；另一方面，研究选择的物种多样性领域不仅种间差异巨大，而且同一物种在不同文本集中的描述也存在差异。因此，本研究的基本思路是在最大限度地识别领域概念的基础上，尽可能完整地抽取描述文本所含信息。由于没有预先定义的模板，从知识共享和支持应用本体构建的角度出发，笔者采用 RDF 模型表示被识别和抽取的信息。RDF 表达式的基本结构是三元组，每个三元组由一个主体、一个谓词和一个客体组成。在本研究中，主体通常是被描述的物种或器官结构(本体中的类或实例)，谓词是其所具有的属性(本体中定义的属性)，客体是属性的值(本体中的类、实例或文字)。

依据已标注语句构建三元组的基本过程是：通过标签匹配判定所描述的特征(谓词)及特征值(客体)；依据客体的类型和谓词确定主体的类型；搜寻与之匹配的标签确定被描述对象(主体)，或依据上下文关系补充被描述对象。为此，笔者利用正则表达式编写了一组规则。

根据适用范围，这些规则被分为通用规则和专用规则两类：

1)通用规则，即针对具有共性的描述形式定义的提取规则。譬如，花瓣、叶、茎的描述往往都会涉及颜色、形态等特征，而且表述形式相近。如："叶片狭长圆形，……，上面深绿色……。花葶直立，长 16-22 厘米，绿色……；花苞片近椭圆形，绿色……"。在标注过程中，"深绿色"、"绿色"都将被赋予标识"ys:ins-color"。通过本体和领域词典可以判定，

"深绿色"、"绿色"是"颜色"类的实例，值域为"颜色"类的属性是"has_color"，"has_color"的定义域为"植物解剖结构"。结合标识"jg"可形成三个三元组："叶片的上面"has_color"深绿色"、"花葶"has_color"绿色"、"花苞片"has_color"绿色"。

2)专用规则，即针对叶、茎、花这三种复合器官的一些特有描述形式而定义的提取规则。譬如，在描述花的语句中出现"'植物解剖结构'+'数量'"的形式，则可推断其为描述花内部结构的数量，抽取时需补充信息并合理设定语序。例如，"退化雄蕊 2"的表示结果为"花"has_part ["退化雄蕊"has_quantity"2"]。

②规则调用逻辑

从性能出发，遵循"从特殊到一般"的准则调用规则，具体过程如图 2 所示：

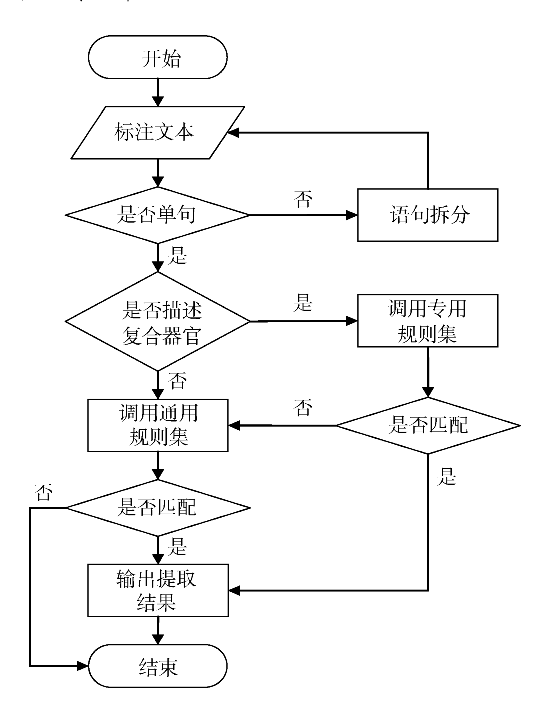


图 2　抽取规则调用逻辑图

③隐性信息识别

除字符所携带的显性信息外，文本还含有一些未通过字符表达的隐性信息。这类隐性信息基本都出现在结构比较复杂的复合器官描述中。譬如，前述"退化雄蕊 2"就是比较典型的例子。类似情况，使用专用规则补充其隐藏的信息。

此外，由于物种多样性描述文本中的句子(以"。"、";"为分隔符)较长，结构复杂，而且常涉及多个描述主体。因此，以子句(以","为分隔符)作为分析和抽取的基本单元。这种处理方式的优点在于显著降低了句子的分析难度，但同时也带来主语(即描述主体)信息缺损的问题。针对这一问题，分两种情况补充描述主体：

1)组合性补足。若当前子句中有表示空间部位的概念(以"kj"标识)，则将前一子句的描述主体和当前子句中表示空间部位的概念拼接，构成当前子句的描述主体。例如，前例中的子句"上面深绿色"。"上面"是表示植物空间部位的概念，与前一子句的描述主体"叶片"拼接，组合成当前子句的描述主体"叶片的上面"。

2)顶替性补足。若当前子句中没有表示植物空间部位的概念，则直接以前一子句的描述主体作为当前子句的描述主体。例如，对于前例中的子句"长 16-22 厘米"，程序将直接补充前一子句的描述主体"花葶"。

④抽取过程示例

为了清晰地呈现规则在抽取过程中的作用方式，以"叶纸质至厚纸质，互生或在短枝顶端簇生，卵状椭圆形或卵状矩圆形，长 3-6 厘米，宽 1.6-3.5 厘米，顶端圆形或钝，常有小尖头，基部圆形或近心形，上面绿色，无毛，下面灰白色，仅脉腋被疏微毛，侧脉每边 8-10 条；" 中的内容为例具体说明。

1)规则的基本作用方式。"叶纸质至厚纸质"----->I.识别出"纸质"和"厚纸质"这两个标为'zd'的特征词，借以质地的定义域寻得前方标记为'jg'的"叶"，同时获取关系"has_texture"，最终确定信息为"叶"has_texture"纸质"、"叶"has_texture"厚纸质"。II.同时保留该句中主语"叶"。

2)复合结构解析及代替性主语补足。"互生或在短枝顶端簇生"----->I.拆分复合句为"互生"、"在短枝顶端簇生"这两句短句。II.通过保留主语对分句缺失成分进行补足，补足后分别为"叶互生"和"叶在短枝顶端簇生"。III.参照上例解析关系，得到"叶"has_arrangement"互生"、"叶"has_arrangement "簇生"。

3)数据属性识别。"长 3-6 厘米"----->I.识别得到标记为'xt'的"长"、标记为'm'的"3-6"、标记为'dw'的"厘米"。II.通过组合匹配判断确认捕捉到内容为"长"的数据属性，以"长"为条件得到关系 has_length，由于标记'dw'的出现还会增加一层为"3-6" has_unit"厘米"的附属关系，并进一步组合获得 has_length["3-6" has_unit"厘米"]。III.通过上一层级的保留主语对缺失主语进行补足，得到"叶"has_length ["3-6" has_unit"厘米"]。

4)复合结构解析、组合性主语补足、程度识别。"基部圆形或近心形"----->I.拆分复合句为"基部圆形"和"基部近心形"。II.通过"基部"的标记'kj'可知需进行组合性补足，从而获得补足后语句"叶基部圆形"和"叶基部近心形"。III.针对"叶基部近心形"中标记为'cd'的"近"与紧邻其后的标记为'xz' 的"心形"，判断调取"has_degree"关系，形成"心形"has_degree"近"。IV.参照通用提取机制，整合附属关系，最终获得"叶基部"has_shape"圆形"和"叶基部"has_shape["心形"has_degree "近"]。

(4) 概念识别

概念系统的完备性是影响信息抽取性能最重要的因素。如果领域本体已经非常完善，那么完全没有必

要执行概念识别过程。而目前，中文植物物种多样性本体虽然已包含 4 000 多条实例，但是不同文本集在描述分类单元模式的选择、所使用术语以及数据表现形式等方面都存在差异，因此，可能还有许多概念未纳入现有领域本体。鉴于此，在概念标注前，运用 CRF 算法检验是否存在未纳入本体的概念。

①特征选择

中文是由独立的字组合成具有特定语义的词，进而依据语法规则组织成句形成文本，词与词之间没有分隔标志。因此，使用 CRF 算法是以字还是词为特征，一直存在分歧。课题组的实验结果表明，以词为特征识别中文植物物种多样性描述文本中的未登录词，其性能优于以字为特征[21]。为了优化识别性能，在词特征的基础上进一步增加词性、词长、相关度、信息熵等特征。

1)词性。提取领域本体中的概念作为用户词典支持 ICTCLAS 分词和词性标注。ICTCLAS 将用户词典所含词条的词性均标注为“un”。对于未包含在用户词典中的字符串，在 ICTCLAS 的切分和词性标注结果的基础上拆分成单字，并赋予所标注的词性标记。例如，ICTCLAS 赋予“主枝”“n”(名词)词性，若其未包含在用户词典中，则拆分为“主”和“枝”，词性均标注为“n”。

2)词长。指词语包含的字数，一般介于 1-5 之间。

3)相关度。相关度反映相邻字之间结合的紧密程度。字串 W 的相关度如下所示：

$$\mathrm{rel(W)} = \frac{n(n_{11} \times n_{12} - n_{12} \times n_{21})^2}{n_{1*}} \times n_{2*} \times n_{*1} \times n_{*2} \qquad (1)$$

若字串 W 的首字为 A，次字为 B，则 n 为语料库所有二元组的串频，即 $n=n_{11}+n_{12}+n_{21}+n_{22}$。其中，$n_{11}$ 为首字为 A 次字为 B 的串频，n_{12} 为首字为 A 次字非 B 的串频，n_{21} 为首字非 A 次字为 B 的串频，n_{22} 为首字非 A 次字非 B 的串频；$n_{1*}=n_{i1}+n_{i2}\ (i=1,2)$; $n_{*j}=n_{1j}+n_{2j}\ (j=1,2)$。

相关度的值是连续值，需要进行离散化处理。在实验的基础上，本研究将其等频率分为 5 个等级，即按照从高到低的顺序，将值最大的 20%赋予“1”，值最小的 20%赋予“5”，以此类推。

4)信息熵。信息熵可以用来界定词与词之间的边界。词 W 的信息熵如下所示：

$$\mathrm{H(W)} = -\sum p\log(p) \qquad (2)$$

其中，p 表示该词左右连接的不同词(字)的概率。

信息熵也是连续值，本研究使用的离散方法是：计算每个字(或词)的左右信息熵，比较大小，若左信息熵大于右信息熵，则将特征标记为“rgh”(right)，说明该字(或词)倾向于与右边的字(或词)链接，左边更可能是词语的边界，否则标记为“lft”(left)。

②标注集

本研究在四位标注集 BIEO 的基础上定义了 SBIEO 标注集。其中，S(Single)表示单字词，B(Begin)表示术语的第一个字，I(In)表示中间的字，E(End)表示最后一个字，O(Out)表示当前字不在术语中。每个 S 或连续的 B(I)E 构成一个术语。例如：字符串“小枝多少密被短伏毛，近方形，粗壮，稍弯曲，散生皮孔”的标注结果如下：小 B/枝 E/多少 O/密 O/被 O/短 O/伏 O/毛 O/，O/近 O/方形 O/，O/粗壮 O/，O/稍 O/弯曲 O/，O/散 O/生 O/皮 B/孔 E/。

③工具与特征模板

调用 CRF++0.58 作为标注工具，以词、词性、词长、相关度、信息熵为特征构建模板。鉴于术语词长一般不超 5，故将窗口长度设为 5。

3　植物物种多样性文本的信息抽取实验

3.1　样　本

中国在线植物志(http://frps.eflora.cn)收录了 301 科 3 408 属 31 142 种植物的科学名称、形态特征、生态环境、地理分布、经济用途和物候期等信息，是最具影响力的中文植物学网络信息源之一[22]。本研究利用自主开发的爬虫程序，从该网站获取鼠李科勾儿茶属、兰科角盘兰属和兜兰属各 17 个物种的描述文档，构成测试集。

以人工方式逐句分析文本内容，建立评价抽取结果的标准答案数据集。该数据集为文档–子句–三元组的映射，共含 4 734 个信息点(三元组)。

3.2　性能评价指标

采用准确率(Precision)和召回率(Recall)评价信息抽取性能，同时引入 F 值作为均衡准确率和召回率的指标。计算公式如下：

$$\mathrm{Precision(P)} = \frac{\text{被准确提取的三元组数量}}{\text{提取出的三元组数量}} \qquad (3)$$

$$\mathrm{Recall(R)} = \frac{\text{被准确提取的三元组数量}}{\text{应提取出的三元组数量}} \qquad (4)$$

$$F = \frac{2 \times P \times R}{P + R} \qquad (5)$$

3.3　实验结果及分析

(1) 整体抽取性能统计

表 2 显示，系统从测试样本中共提取 4 697 条信息，整体准确率、召回率分别约为 0.86、0.85，F 值为 0.85。准确率和召回率表现均衡，抽取性能较理想。

表 2　抽取性能统计汇总

描述文档	提取数	正确数	遗漏数	准确率	召回率	F 值
鼠李科勾儿茶属	1 108	1 002	107	0.904332	0.903517	0.903924
兰科兜兰属	1 773	1 472	436	0.830231	0.771488	0.799783
兰科角盘兰属	1 816	1 548	169	0.852423	0.901573	0.876309
总计	4 697	4 022	712	0.856291	0.849599	0.852932

石静以《中国高等植物图鉴》中的 60 种植物的描述文本作为测试样本，涉及旋花科、茄科、杜鹃花科等 12 个科。特征描述信息抽取的平均准确率和召回率分别为 0.868、0.7138，F 值为 0.7834[18]。应注意，本研究是将依据抽取结果构建的三元组与标准答案比对，计算准确率和召回率；而石静的研究则是依据模板填充结果计算性能指标。两者采用的测试样本、计算依据都不相同，在理论上不宜直接比较两者的性能差异。

(2) 科属间的差异分析

表2中的数据显示，系统抽取不同科、属描述文档中的信息，性能可能存在差异。利用SPSS软件比较鼠李科与兰科、兰科的兜兰属与角盘兰属描述文档中信息的抽取性能(F值的均值)，分析结果表明组间的确存在差异。为寻找差异产生的原因，按描述主体重新对三组文档信息的抽取性能进行统计，如图3至图5所示。从图4与图3、图5的比较可知，兜兰属样本中茎和根的描述信息抽取效果不佳是导致评价指标偏低的主要原因。当然，这并不意味着所有物种茎和根描述信息的抽取效果一定不理想，图5很好地说明了这一点。

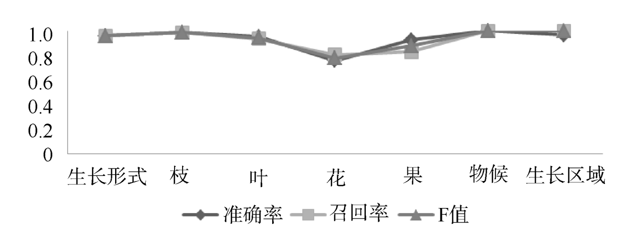


图 3　鼠李科勾儿茶属样本中各描述主体信息的抽取性能

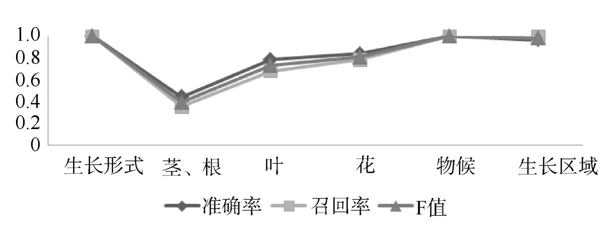


图 4　兰科兜兰属样本中各描述主体信息的抽取性能

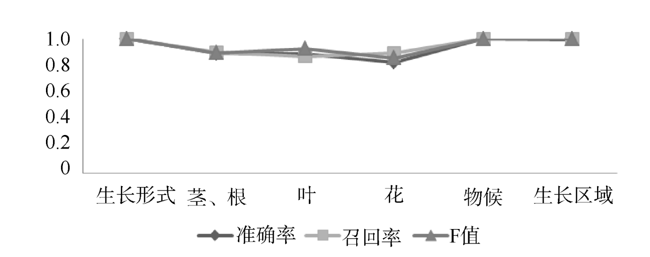


图 5　兰科角盘兰属样本中各描述主体信息的抽取性能

(3) 文档间的差异分析

更进一步地，笔者希望明确科属间的抽取性能差异是源于少数极端样本的影响，还是组间样本整体的差异所导致。为此，对三组样本中的文档分别随机分配 1-17 的序号，比较每篇文档中信息的抽取性能。从图 6、图 7 能够清楚地看到，类属相同的物种，其描述文本信息抽取的准确率、召回率虽有波动，但总体比较平稳。这表明导致科属间性能差异的主要原因并非来自个例的影响，这一点在图 7 中表现尤为清晰。

图 6、图 7 同时显示，兜兰属编号为 4、8、13 的文献抽取结果的召回率、准确率都较低。为此，分析这三篇文献中对茎和根的描述语句，发现错误原因在于对“有少数稍肉质而被毛的纤维根”的解析。该句描述的是附属结构“纤维根”，但是因“而”这一关联词，使该句在处理时被拆分为两个单句，破坏了原有语义，导致描述主体判断错误。

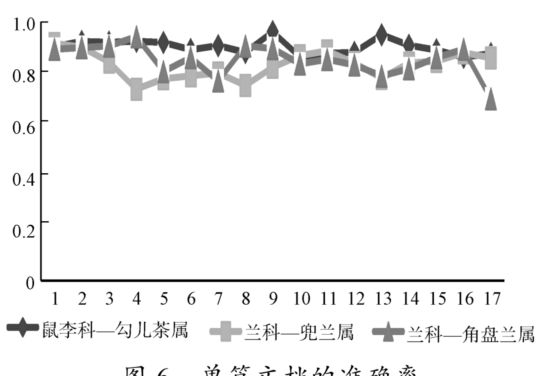


图 6　单篇文档的准确率

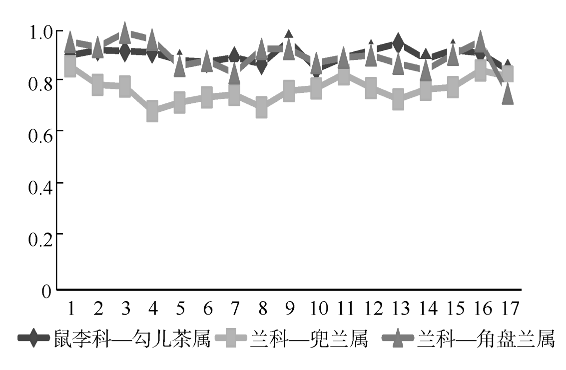


图 7　单篇文档的召回率

(4) 存在的问题

从实验结果来看，本研究获得了比较理想的抽取结果。但归纳错抽和漏抽的实例，发现仍有一些问题有待解决。

①以比较或排除方式描述。譬如，“较中裂片长很多或稍较长”、“雄蕊长于花瓣”、“除背面中脉近基部处具长柔毛外余均无毛”。

②与结构部位关联的生长趋势描述。譬如，“中部以上向先端渐狭”、“向末略变狭”、“从蕊喙下向外伸出”。

③具有多项特征值的描述。譬如，“侧脉每边 7-13 条通常 9-10 条”、“顶端钝或圆形稀短渐尖”、“叶(1-)2 枚极罕为 3 枚”。

4　结　语

本研究设计并实现了一个中文植物物种多样性描述文本信息抽取方案，性能(F 值)达到 0.85。方案的设计思路兼顾适应性和性能。以本体为支撑，采取语段、语句、概念逐级筛选和标注的策略，依据规则实现描述文本中信息的抽取。在理论上，该方案建立的框架能支持生物物种多样性、病症乃至商品等多种描述性文本中信息的抽取。在应用方面，本研究不仅开发了一套实用的信息抽取系统，还建立了一个较完善的植物物种多样性领域本体，同时提出了一个比较成熟的植物物种多样性领域概念识别方法。当然，研究还可以进一步修正和完善。譬如，以上所提及影响抽取性能的三个问题，以及如何组织构建的三元组集合，使其准确地表示原文语义。

(致谢：感谢中国科学院植物研究所文献与信息中心刘凤红高级工程师、南京林业大学陈金慧教授在本体构建过程中给予的支持。)

参考文献：

[1] BHL. Biodiversity Heritage Library [EB/OL]. [2015-09-27]. http://www.biodiversitylibrary.org/.

[2] Thessen A E, Cui H, Mozzherin D. Applications of Natural Language Processing in Biodiversity Science [J]. Advances in Bioinformatics, 2012: Article ID 391574. doi: 10.1155/2012/391574.

[3] Vanel J M. Worldwide Botanical Knowledge Base [EB/OL]. [2011-10-11]. http://wwbota.free.fr/.

[4] 郑家恒，菅小艳. 农作物信息抽取系统的设计与实现[J]. 计算机工程, 2006, 32(7): 197-198, 220. (Zheng Jiaheng, Jian Xiaoyan. Design and Realization of the System of Farm Crop Information Extraction [J]. Computer Engineering, 2006, 32(7): 197-198, 220.)

[5] Cui H, Heidorn P. The Reusability of Induced Knowledge for Automatic Semantic Markup of Taxonomic Descriptions [J]. Journal of the American Society for Information Science and Technology. 2007, 58(1): 133-149.

[6] 段宇锋，黑珍珍，鞠菲，等. 基于自主学习规则的中文物种描述文本的语义标注研究[J]. 现代图书情报技术, 2012(5): 41-47. (Duan Yufeng, Hei Zhenzhen, Ju Fei, et al. Study on Semantic Markup of Species Description Text in Chinese Based on Auto-learning Rules [J]. New Technology of Library and Information Service, 2012(5): 41-47.)

[7] 段宇锋，黑珍珍，鞠菲，等. 基于贝叶斯分类的中文物种描述文本的语义标注研究[J]. 情报学报, 2012, 31(8): 805-812. (Duan Yufeng, Hei Zhenzhen, Ju Fei, et al. Semantic Annotation of Species Description Text in Chinese Literature by Naïve Bayes Classifier [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(8): 805-812.)

[8] 段宇锋，朱雯晶，陈巧，等. 朴素贝叶斯算法与 Bootstrapping 方法相结合的中文物种描述文本语义标注研究[J]. 现代图书情报技术, 2014(5): 83-89. (Duan Yufeng, Zhu Wenjing, Chen Qiao, et al. Semantic Annotation of Species Description Text in Chinese by Combining Naïve Bayes Algorithm with Bootstrapping Method [J]. New Technology of Library and Information Service, 2014(5): 83-89.)

[9] Taylor A. Extracting Knowledge from Biological Descriptions [C]. In: Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases. 1995: 114-119.

[10] Wood M M, Lydon S J, Tablan V, et al. Using Parallel Texts to Improve Recall in IE [C]. In: Proceedings of Recent Advances in Natural Language Processing (RANLP'03). 2003: 505-512.

[11] Tang X, Heidorn P B. Using Automatically Extracted

Information in Species Page Retrieval [OL]. [2011-08-10]. http://www.tdwg.org/proceedings/article/view/195/.
[12] Soderland S. Learning Information Extraction Rules for Semi-Structured and Free Text [J]. Machine Learning, 1999, 34(1-3): 233-272.
[13] Abascal R, Sanchez J A. X-tract: Structure Extraction from Botanical Textual Descriptions [C]. In: Proceeding of the String Processing & Information Retrieval Symposium & International Workshop on Groupware.1999: 2-7.
[14] Diederich J, Frotuner R, Milton J. Computer-assisted Data Extraction from the Taxonomical Literature [OL]. [2011-08-15]. http://math.ucdavis.edu/~milton/genisys.html.
[15] Cui H. CharaParser for Fine-grained Semantic Annotation of Organism Morphological Descriptions [J]. Journal of the American Society for Information Science and Technology, 2012, 63(4): 738-754.
[16] Cui H, Singaram S, Janning A. Combine Unsupervised Learning and Heuristic Rules to Annotate Morphological Characters [J]. Proceedings of the American Society for Information Science and Technology, 2011, 48(1): 1-9.
[17] 沙丽华. 面向领域文档的语义标注方法研究[D]. 长春: 吉林大学, 2009. (Sha Lihua. Research on Semantic Annotation for Domain Documents [D]. Changchun: Jilin University, 2009.)
[18] 石静. 基于本体的植物信息抽取与分析研究[D]. 杨凌: 西北农林科技大学, 2010. (Shi Jing. Information Extraction and Analysis Based on Plant Ontology [D]. Yangling: Northwest Agriculture and Foresty University, 2010.)
[19] Gruber T R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing [J]. International Journal of Human-Computer Studies, 1995, 43(5-6): 907-928.
[20] 向阳, 王敏, 马强. 基于 Jena 的本体构建方法研究[J]. 计算机工程, 2007, 33(14): 59-61. (Xiang Yang, Wang Min, Ma Qiang. Research on Jena-based Ontology Building [J]. Computer Engineering, 2007, 33(14): 59-61.)
[21] 段宇锋, 朱雯晶, 陈巧, 等. 条件随机场与领域本体元素集相结合的未登录词识别研究[J]. 现代图书情报技术, 2015(4): 41-49. (Duan Yufeng, Zhu Wenjing, Chen Qiao, et al. The Study on Out-of-Vocabulary Identification on a Model Based on the Combination of CRFs and Domain Ontology Elements Set [J]. New Technology of Library and Information Service, 2015(4): 41-49.)
[22] 中国植物志编辑委员会. 中国植物志[DB/OL]. [2007-09-28]. http://frps.eflora.cn/. (Flora of China Editorial Committee. Flora of China [DB/OL]. [2007-09-28]. http://frps.eflora.cn/.)

作者贡献声明:

段宇锋: 提出研究思路, 设计研究方案, 论文起草和修订;
黄思思: 开发程序, 采集、清洗和分析数据。

收稿日期: 2015-09-14
收修改稿日期: 2015-09-28

Information Extraction from Chinese Plant Species Diversity Description Text

Duan Yufeng Huang Sisi
(Business School, East China Normal University, Shanghai 200241, China)

Abstract: [**Objective**] To extract information from Chinese plant species diversity description text. [**Methods**] Take the plant species diversity domain ontology as the foundation, and adopt the strategy of stepwise selection and annotation on paragraph, sentence and concept. [**Results**] A sample including 4 734 information points is used to test. The value of extraction accuracy rate, recall rate and F-measure achieves 0.86, 0.85 and 0.85 respectively. [**Limitations**] In order to solve the problems on extracting information from description text, the rule set should be improved in the future. [**Conclusions**] The research scheme can fulfill the information extraction from Chinese plant species diversity description text effectively.

Keywords: Information extraction Plant species diversity description text Chinese information processing Ontology